UTMD Working Paper
The University of Tokyo
Market Design Center

UTMD-013

Epistemological Implementation of Social Choice Functions

Hitoshi Matsushima
University of Tokyo

First Version: July 20, 2021
This Version: October 26, 2022

Games and Economic Behavior 136 (2022) 389–402

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Epistemological implementation of social choice functions ☆

Hitoshi Matsushima

Department of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

ARTICLE INFO

Article history:
Received 27 July 2021
Available online 20 October 2022

JEL classification:
C72
D71
D78
H41

Keywords:
Unique implementation
Weak honesty
Common knowledge on selfishness
Ethical social choice function
Quadratic scoring rule

ABSTRACT

We investigate the implementation of social choice functions (SCFs) from an epistemological perspective. We consider the possibility that in higher-order beliefs there exists an honest agent who is motivated by intrinsic preference for honesty as well as material interest. We assume weak honesty, in that, although any honest agent has a cost of lying that is positive but close to zero, she (or he) is mostly motivated by material interests and even tells white lies. This study assumes that all agents are fully informed of the physical state, but "all agents are selfish" never happens to be common knowledge in epistemology. We show the following positive results for the implementability: with three or more agents, any SCF is uniquely implementable in the Bayesian Nash equilibrium (BNE). An SCF, whether material or nonmaterial (ethical), can be implemented even if all agents are selfish and "all agents are selfish" is mutual knowledge.

1. Introduction

This study investigates the implementation problem of social choice functions (SCFs) from an epistemological perspective. A central planner attempts to implement the desirable allocation implied by an SCF contingent on the state. She (or he)[1] does not know the (physical) state, while there exist multiple agents (participants) who are fully informed of the state. The central planner would like to hear from these informed agents about what the correct state is, but they might lie and manipulate the information to make the central planner's decisions more beneficial for them. Under these circumstances, the central planner attempts to design a decentralized mechanism, which consists of message spaces, an allocation rule, and a payment rule, and incentivize these agents to announce about the state sincerely in this mechanism. The question is, under what condition can the central planner implement the SCF as a unique equilibrium outcome?

We assume that each agent is either selfish or honest. A selfish agent is only concerned about her material utility, while an honest agent is concerned about the intrinsic preference for honesty as well. However, importantly, we do not assume any possibility that there exists an honest agent as a participant in the central planner's problem. Instead, we consider the epistemological possibility that an honest agent exists, not in the mechanism, but in the participants' higher-order beliefs. Hence, we assume incomplete information concerning the agents' epistemological types, while we assume complete

☆ This study was supported by a grant-in-aid for scientific research (KAKENHI 20H00070) from the Japan Society for the Promotion of Science (JSPS) and the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT). Matsushima (2021d) is the earlier version of this study, which includes an extension of this study to incomplete information concerning the state. I am grateful to Professor Shunya Noda for his useful comments and discussions. I am also grateful to the editor, the advisory editor, and the referees of this journal for their helpful comments and suggestions. All errors are mine.

E-mail address: hitoshi@e.u-tokyo.ac.jp.

[1] For convenience, this study uses the gendered pronoun "she," instead of "she or he" or "they."

https://doi.org/10.1016/j.geb.2022.10.007

information concerning the physical state. By considering the epistemological type space in this manner, we show that a slight possibility of an honest agent in higher-order beliefs incentivizes all agents, whether selfish or honest, to behave sincerely.

Importantly, we assume "no common knowledge of selfishness" (NCKS) in the sense that "all agents are selfish (i.e., all agents are motivated only by their monetary interests)" never happens to be common knowledge. With this, we demonstrate the following positive results for the implementability: with three or more agents, any SCF is uniquely implementable in the Bayesian Nash equilibrium (BNE).

We require each agent to announce probability distributions over states. We then utilize a simple form of the quadratic scoring rule (Brier, 1950), which aligns agents' payoffs with the distance between their messages, as a part of the payment rule design. The quadratic scoring rule plays a significant role in incentivizing all agents, whether selfish or honest, to announce sincerely as unique BNE behavior, whenever the common knowledge of selfishness is eliminated.

This study requires only a weak honesty, where an honest agent has a cost of lying that is positive but close to zero. Hence, the influences of the intrinsic preferences for honesty on decision making are arbitrarily small. Agents do not expect the possibility that there exists an honest participant; that is, they may have mutual knowledge that all agents are selfish (i.e., all agents know that all agents are selfish). Despite these weaknesses in honesty, the central planner can elicit correct information from all agents if "all agents are selfish" never happens to be common knowledge, that is, under NCKS.

2. Literature review

The early literature on social choice and implementation theory has assumed that "all agents are selfish" is common knowledge, and focused on those considerations of SCFs that are material, that is, those that depend only on agents' material utilities (Arrow, 1951; Hurwicz, 1972; Gibbard, 1973; Satterthwaite, 1975; Maskin, 1977/1999; Abreu and Matsushima, 1992a, 1992b).[2] With this common knowledge, it is impossible in principle to implement any SCF that is nonmaterial, or ethical; that is, it depends not only on an agent's material utilities but also on nonmaterial factors such as ethics, equity, fairness, future generation, and environmental concerns. This study suggests a highly positive potential for implementing such nonmaterial, or ethical, SCFs.[3]

Matsushima (2008a) and Matsushima (2008b) are the pioneering works incorporating an intrinsic preference for honesty into implementation theory and demonstrating a new research trend to overcome the above-mentioned difficulty. Matsushima (2008a) showed that in complete information environments concerning the state, with three (or more) agents, any SCF, whether material or nonmaterial, is uniquely implementable if there exists a weakly honest agent who dislikes telling a white lie. Matsushima (2008b) showed that in asymmetric information environments concerning the state, any incentive-compatible SCF, whether material or nonmaterial, is uniquely implementable if all agents are honest. Many subsequent studies such as Dutta and Sen (2012), Kartik et al. (2014), Saporiti (2014), Ortner (2015), and Mukherjee et al. (2017) have showed their respective positive results.[4]

This study makes two significant advances in this line of research under complete information concerning the state. First, the previous works assumed that there exists an honest agent as a participant in the mechanism, at least with a positive probability, while this study does not assume it at all. We only rule out the case in which "all agents are selfish" is common knowledge: we permit the case where all agents are selfish and "all agents are selfish" is mutual knowledge.

Second, previous works assumed that an agent never tells a white lie, that is, a lie that does not influence material utilities. We can regard this assumption as one of the weakest conditions for an agent to be honest. However, real people may not be honest even under such minimum conditions. They may be influenced by various adversarial motives to tell white lies. Even selfish people may have nonmaterial motives, whether prosocial or adversarial, as lexicographical preferences. Importantly, it is generally difficult to identify whether an agent is the one who never tells white lies. In this sense, we contend that this assumption is restrictive. In contrast to this assumption, this study permits agents, whether selfish or honest, to tell white lies.

This study excludes the possibility that an agent behaves dishonestly beyond her material interests. Many empirical and experimental studies indicate that human beings are not purely motivated by monetary payoffs but have intrinsic preferences for honesty. Abeler et al. (2019) provided a detailed meta-analysis using data from 90 studies involving more than 44,000 subjects across 47 countries, showing that subjects who were in trade-offs between material interest and honesty gave up a large fraction of potential benefits from lying.[5] These empirical and experimental studies support the validity of this study's assumptions.

We permit that agents can tell white lies. We do not assume that there exists an honest agent in the mechanism. To make full use of the slight possibility of weak honesty in epistemology, we utilize the quadratic scoring rules, which can set

[2] For surveys of implementation theory, see Moore (1992), Jackson (2001), and Maskin and Sjöström (2002).

[3] An exception is Matsushima (2019, 2021a), which assumed that the state is ex-post verifiable and proved that any SCF, whether material or nonmaterial, is uniquely implementable even if "all agents are selfish" is common knowledge.

[4] See also Matsushima (2013), Yadav (2016), Lombardi and Yoshihara (2017, 2018, 2019), Dogan (2017), and Savva (2018).

[5] Various works in behavioral economics and decision theory have modeled preferences for honesty, such as a cost of lying (e.g., Ellingsen and Johannesson, 2004; Kartik, 2009; Kartik et al., 2007), a reputational cost (e.g., Mazar et al., 2008), guilt aversion (e.g., Charness and Dufwenberg, 2006), and a cost of false evidence submission (Kartik and Tercieux, 2012).

aside various non-selfish motives and prioritize the agent's monetary interest. As Abeler, Nosenzo, and Raymond point out, the intrinsic preference for honesty remains included, and an honest agent prefers announcing more honestly than selfish agents. This will be the driving force for a tail-chasing competition through which each agent announces more honestly than the other agents, reaching a point at which all agents report honestly.

The quadratic scoring rule is one of the standard mechanism design methods for partial implementation.[6] However, if "all agents are selfish" is common knowledge, there exists a serious multiplicity of unwanted BNEs: "all agents tell the same lie" is a BNE, regardless of what this lie is.

Matsushima and Noda (2020) first found in the context of information elicitations that truth-telling is a unique BNE if there exists an honest agent, thereby suggesting that the quadratic scoring rule design is a potentially powerful solution not only for partial implementation but also for unique implementation. This study generalizes the usefulness of the quadratic scoring rule design by showing positive results for the general implementability of SCFs.

The equilibrium analysis of games with behavioral agents and asymmetric information has a long history. Kreps et al. (1982) studied how the existence of behavioral agents changes the equilibria of finitely repeated games. Postlewaite and Vives (1987), Carlsson and van Damme (1993), and Morris and Shin (1998) studied how incomplete information shrinks the set of equilibria. These studies focused on the analysis of given games, while our focus is on the design of mechanisms that can take full advantage of the potential existence of behavioral agents in higher-order beliefs.

Similar to Rubinstein's (1989) email game, this study contrasts the outcome under common knowledge and "almost common knowledge." Rubinstein's (1989) email game demonstrates that "almost common knowledge" could lead to an unintuitive outcome; while our study demonstrates the vulnerability of the common knowledge assumption, its implication contrasts Rubinstein's. The intuitive outcome is truth-telling, and people can naturally expect that a truthful strategy profile is a focal point, while there are many equilibria under common knowledge of selfishness. By carefully designing a "game" (mechanism), we can eliminate all the unwanted and unintuitive equilibria where "all agents are selfish" is not common knowledge (while it could be "almost common knowledge").

The remainder of this paper is organized as follows. The basic model is presented in Section 3. In Section 4, we semantically define a class of indirect mechanisms and then define the intrinsic preference for honesty. We define the unique implementation in BNE and state that any SCF is uniquely implementable if "all agents are selfish" never happens to be common knowledge. Section 5 concludes.

3. The model

This study investigates a situation in which a central planner attempts to achieve a desirable allocation contingent on the state. Let $N \equiv \{1, ..., n\}$ denote the finite set of all agents, where $n \geq 2$. Let A denote the non-empty and finite set of all the allocations. Let Ω denote the non-empty and finite set of (physical) states. The *social choice function* (SCF) is defined as $f : \Omega \to \Delta(A)$.[7] For every $\omega \in \Omega$, $f(\omega) \in \Delta(A)$ implies the desirable distribution of allocation at state ω.[8] We assume that the central planner does not know the state, while all agents are fully informed of the state.

Each agent is either *selfish* or *honest*. Details on the meaning of selfishness and honesty will be explained in Section 4. No agent knows if the other agents are selfish or honest. To describe agents' higher-order beliefs concerning their selfishness and honesty, we define an *epistemological type space* as follows:

$$\Gamma \equiv (T_i, \pi_i, \theta_i)_{i \in N},$$

where $t_i \in T_i$ is agent $i's$ epistemological type, $\pi_i : T_i \to \Delta(T_{-i})$, and $\theta_i : T_i \to \{0, 1\}$.[9] Agent i is selfish (honest) if $\theta_i(t_i) = 0$ ($\theta_i(t_i) = 1$, respectively). Agent i expects that the epistemological types of other agents are distributed according to the probability measure $\pi_i(t_i) \in \Delta(T_{-i})$. We assume that there exists a common prior $\pi \in \Delta(T)$ from which $(\pi_i)_{i \in N}$ is derived. We assume that the state ω and the epistemological type profile $t = (t_i)_{i \in N}$ is independently drawn.

We call a subset of epistemological type profiles $E \subset T \equiv \times_{i \in N} T_i$ an event. For convenience, for each event $E \subset T$, we write $\pi_i(E|t_i) = \pi_i(E_{-i}(t_i)|t_i)$, where we denote $E_{-i}(t_i) \equiv \{t_{-i} \in T_{-i} | (t_i, t_{-i}) \in E\}$. We denote by $E^* \subset T$ the event that all agents are selfish, that is,

$$E^* \equiv \{t \in T | \forall i \in N : \theta_i(t_i) = 0\}.$$

For each agent $i \in N$, we define the set of all selfish types as

[6] See Cooke (1991) for a survey of scoring rules. For the applications to mechanism design, see for example Johnson et al. (1990), Matsushima (1990, 1991, 1993, 2007), Aoyagi (1998), and Miller et al. (2007). A number of studies extended the scoring rule of Brier (1950) to a setting in which a central planner collects information from a group of agents (e.g., Dasgupta and Ghosh, 2013; Prelec, 2004; Miller et al., 2005; Kong and Schoenebeck, 2019). Previous studies commonly assumed that all agents are selfish and, thus, suffered from the multiplicity of equilibria in a "single-question" setting in which the state is realized only once (as in our model).

[7] We denote by $\Delta(Z)$ the space of probability measures on the Borel field of a measurable space Z. If Z is finite and $\rho \in \Delta(Z)$ satisfies $\rho(z) = 1$ for some $z \in Z$, I will simply write $\rho = z$.

[8] This study considers both deterministic and stochastic SCFs.

[9] We denote $Z \equiv \times_{i \in N} Z_i$, $Z_{-i} \equiv \times_{j \neq i} Z_j$, $z = (z_i)_{i \in N} \in Z$, and $z_{-i} = (z_j)_{j \neq i} \in Z_{-i}$.

$$E_i^* \equiv \{t_i \in T_i | \theta_i(t_i) = 0\}.$$

Consider an arbitrary event $E \subset T$. Let

$$V_i^1(E) = \{t_i \in T_i | \pi_i(E|t_i) = 1\},$$

which denotes the set of agent $i's$ types who know the occurrence of E. Let

$$V_i^2(E) = \{t_i \in T_i | \pi_i(\underset{j \in N}{\times} V_j^1(E)|t_i) = 1\},$$

which denotes the set of agent $i's$ types who know the occurrence of $\times_{j \in N} V_j^1(E)$, that is, know that all agents know the occurrence of E. Recursively, for each positive integer $h \geq 3$, let

$$V_i^h(E) = \{t_i \in T_i | \pi_i(\underset{j \in N}{\times} V_j^{h-1}(E)|t_i) = 1\},$$

which denotes the set of agent $i's$ types who know the occurrence of $\times_{j \in N} V_j^{h-1}(E)$, that is, know that all agents know the occurrence of $\times_{j \in N} V_j^{h-2}(E)$. We then define

$$V_i^\infty(E) \equiv \cap_{h=1}^\infty V_i^h(E).$$

An event $E \subset T$ is said to be *common knowledge* at an epistemological type profile $t \in T$ if

$$t \in \underset{i \in N}{\times} V_i^\infty(E).$$

Note that if E is common knowledge at $t \in T$, then

$$\pi_i \left(\underset{j \in N}{\times} V_j^\infty(E) \middle| t_i \right) = 1 \text{ for all } i \in N.$$

It will be explained throughout this study that whether the event of "all agents are selfish" (i.e., $E = E^*$) is common knowledge has a decisive impact on the implementability of a SCF.

Fix an arbitrary positive real number $\varepsilon > 0$. We define a mechanism as $G \equiv (M, g, x)$, where $M = \times_{i \in N} M_i$, M_i denotes agent $i's$ message space, $g : M \to \Delta(A)$ denotes an allocation rule, $x = (x_i)_{i \in N}$ denotes a payment rule, and $x_i : M \to [-\varepsilon, \varepsilon]$ denotes the payment rule for agent i. Here, $\varepsilon > 0$ implies the level of limited liability. Each agent i simultaneously announces a message $m_i \in M_i$, and the central planner determines the allocation according to $g(m) \in \Delta(A)$ and pays the monetary transfer $x_i(m) \in R$ to each agent i.

Each agent $i's$ material benefit is given by a quasi-linear utility $v_i(a, \omega) + r_i$, provided that the central planner determines the allocation $a \in A$ and gives the monetary transfer $r_i \in R$ to agent i at state $\omega \in \Omega$. We assume expected utility for convenience, and denote $v_i(\alpha, \omega)$ the expected payoff derived from stochastic allocation $\alpha \in \Delta(A)$. When all agents announce $m \in M$ in the mechanism G, the resultant expected material payoff is given by $v_i(g(m), \omega) + x_i(m)$.

This study considers a small liability ε, which is positive but close to zero. Quasi-linearity is a convenient, but rather redundant, assumption: all we need for this study is that an agent's material benefit increases as the monetary transfer to her increases. The expected utility assumption is also redundant: see Matsushima (2019, 2021a) for this detail.

4. Unique implementation

We assume $n \geq 3$. From a semantic point of view, we focus on the following class of indirect mechanisms. We fix an arbitrary positive integer $K \geq 1$, the specification of which is explained in Subsection 4.2. Let

$$M_i = \underset{k=1}{\overset{K}{\times}} M_i^k,$$

and

$$M_i^k \subset \Delta(\Omega) \text{ for all } k \in \{1, ..., K\},$$

where we denote $m_i = (m_i^k)_{k=1}^K$, and $m_i^k \in M_i^k$ for each $k \in \{1, ..., K\}$. Each agent i reports K sub-messages at once, which typically concern which state occurs. At each k-th sub-message, agent i announces, not a single state, but a probability distribution over states $m_i^k \in \Delta(\Omega)$. At each of agent $i's$ sub-messages, we assume agent i announces more truthfully, as her announcement at this sub-message grants greater probability to the true state. An agent can announce different distributions across sub-messages. This specification serves to evoke an ethical motive to tell the truth from an agent in a manner such that the agent feels guilty about telling a lie that generates more material benefit that derives directly from the central planner's decision. Further details on the implication of ethical motive and material benefit will be provided shortly.

This study specifies M_i^k as either the set of all distributions over states or the set of all degenerate distributions over states. If M_i^k is specified as the set of all degenerate distributions, we simply write $M_i^k = \Omega$.

Denote $m_i^k = (m_i^k(\omega))_{\omega\in\Omega} \in \Delta(\Omega)$. At each k-th sub-message, agent i announces that each state $\omega \in \Omega$ occurs with a probability of $m_i^k(\omega) \in [0,1]$. We simply write $m_i^k = \omega$ if $m_i^k(\omega) = 1$. We also denote $m_i(\omega) = (m_i^k(\omega))_{k=1}^K \in [0,1]^K$. Importantly, we will consider agent i acting more honestly at state ω when she announces m_i rather than $\tilde{m}_i$, if the vector m_i assigns higher probability to the true state than the vector $\tilde{m}_i$ in each component, that is,[10]

$$m_i(\omega) \neq \tilde{m}_i(\omega) \text{ and } m_i(\omega) \geq \tilde{m}_i(\omega).$$

We define a strategy for agent i as

$$s_i : \Omega \times T_i \to M_i,$$

according to which, agent i with epistemological type t_i announces $m_i = s_i(\omega, t_i) \in M_i$ at the state ω. Denote $s_i = (s_i^k)_{k=1}^K$, $s_i^k : \Omega \times T_i \to M_i^k$, and $s_i(\omega, t_i) = (s_i^k(\omega, t_i))_{k=1}^K$, where $s_i^k(\omega, t_i) \in M_i^k \subset \Delta(\Omega)$ denotes agent $i's$ k-th sub-message. We also denote $s_i(\omega, t_i)(\omega') = (s_i^k(\omega, t_i)(\omega'))_{k=1}^K \in [0,1]^K$, where, at a state ω, agent i with epistemological type t_i announces as her k-th sub-message that each state $\omega' \in \Omega$ occurs with a probability of $s_i^k(\omega, t_i)(\omega') \in [0,1]$. We simply write $s_i^k(\omega, t_i) = \omega'$ if $s_i^k(\omega, t_i)(\omega') = 1$.

We define the *sincere strategy* for agent i, denoted by $s_i^* = (s_i^{*k})_{k=1}^K$, as

$$s_i^{*k}(\omega, t_i) = \omega \text{ for all } i \in N,\ \omega \in \Omega,\ t_i \in T_i, \text{ and } k \in \{1, ..., K\},$$

according to which, agent i announces the state truthfully at any sub-message.

Each agent $i \in N$ is either selfish ($\theta_i(t_i) = 0$) or honest ($\theta_i(t_i) = 1$). If agent i is selfish, she is only concerned with the material benefit; that is, she maximizes the expected value of material benefit:

$$[\theta_i(t_i) = 0]$$
$$\Rightarrow [s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[v_i(g(m), \omega) + x_i(m) | \omega, t_i, s_{-i}]],$$

where we assumed that the other agents announce according to $s_{-i} = (s_j)_{j\neq i}$.[11]

If agent i is honest, she is motivated not only by material benefit but also by an *intrinsic preference for honesty*, depending on the specification of the mechanism in a semantical viewpoint: she has a psychological cost of lying $c_i(m, \omega, t_i, G) \in R$ such that for every $\omega \in \Omega$, $m \in M$, and $\tilde{m}_i \in M_i$,

$$[\theta_i(t_i) = 1,\ m_i(\omega) \neq \tilde{m}_i(\omega),\ m_i(\omega) \geq \tilde{m}_i(\omega), \text{ and} \tag{1}$$
$$v_i(g(\tilde{m}_i, m_{-i}), \omega) + x_i(\tilde{m}_i, m_{-i}) > v_i(g(m), \omega) + x_i(m)]$$
$$\Rightarrow [c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G) > c_i(m, \omega, t_i, G)],$$

and

$$[\theta_i(t_i) = 1 \text{ and } m_i(\omega) = \tilde{m}_i(\omega)] \tag{2}$$
$$\Rightarrow [c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G) = c_i(m, \omega, t_i, G)].$$

From (1), an honest agent feels more guilty (i.e., $c_i(\tilde{m}_i, m_{-i}, \omega, t_i, G) > c_i(m, \omega, t_i, G)$) if she gains material payoffs from telling a more lie (i.e., $m_i(\omega) \neq \tilde{m}_i(\omega)$ and $m_i(\omega) \geq \tilde{m}_i(\omega)$). We do not impose any restriction on the strength of the psychological cost. She maximizes the expected value of the material benefit minus the psychological cost:

$$[\theta_i(t_i) = 1] \Rightarrow [s_i(\omega, t_i) \in \arg\max_{m_i \in M_i} E[v_i(g(m), \omega) + x_i(m)$$
$$- c_i(m, \omega, t_i, G) | \omega, t_i, s_{-i}]].$$

Hence, due to this psychological cost, an agent is willing to act more honestly (more sincerely) if she is honest rather than selfish.

Formally, each agent $i's$ payoff function, $U_i(\cdot; \omega, t_i) : M \to R$, is defined as

$$U_i(m; \omega, t_i) = v_i(g(m), \omega) + x_i(m) \quad \text{if } \theta_i(t_i) = 0,$$

[10] For two vectors $z = (z^k)_{k=1}^K$ and $z = (z^k)_{k=1}^K$, we write $z \geq \tilde{z}$ if and only if $z^k \geq \tilde{z}^k$ for all $k \in \{1, ..., K\}$.

[11] $E[\cdot|\xi]$ denotes the expectation operator conditional on ξ.

and

$$U_i(m;\omega,t_i)=v_i(g(m),\omega)+x_i(m)-c_i(m,\omega,t_i,G)\quad \text{if } \theta_i(t_i)=1.$$

A strategy profile s is said to be a *Bayesian Nash equilibrium* (BNE) in the game associated with the mechanism G if for every $\omega\in\Omega$, $i\in N$, $t_i\in T_i$, and $m_i\in M_i$,

$$E[U_i(s_i(\omega,t_i),m_{-i};\omega,t_i,G)|\omega,t_i,s_{-i}]$$
$$\geq E[U_i(m_i,m_{-i};\omega,t_i,G)|\omega,t_i,s_{-i}].$$

A mechanism G is said to *uniquely implement* an SCF f if there exists the unique BNE s, and it induces the value of f; that is,

$$g(s(\omega,t))=f(\omega)\text{ for all }\omega\in\Omega\text{ and }t\in T,$$

where we denote $s(\omega,t)=(s_i(\omega,t_i))_{i\in N}$. An SCF is said to be *uniquely implementable* if there exists a mechanism that uniquely implements it.

Theorem 1. *Any SCF f is uniquely implementable if*

$$\underset{i\in N}{\times} V_i^{\infty}(E^*)=\phi. \tag{3}$$

Equality (3) implies "no common knowledge of selfishness" (shortly, NCKS), in that "all agents are selfish" never happens to be common knowledge. Theorem 1 states that under NCKS, every SCF is uniquely implementable. The proof of Theorem 1 is presented in the subsequent subsections.

4.1. Special case: information elicitation

To understand the proof of Theorem 1, it is helpful to investigate the information elicitation problem as a special case where we assume that each agent's material payoff is irrelevant to the allocation; that is,

$$v_i(a,\omega)=0\text{ for all }i\in N,\ a\in A,\ \text{and }\omega\in\Omega.$$

Note that due to this irrelevancy, if an agent never tells a white lie, the central planner can get the true information out of her by simply asking her about which state occurs. However, since this study allows agents to tell white lies, we need to carefully design incentive devices for information elicitations by taking advantage of the epistemological possibility of honesty. The following proposition states that NCKS guarantees unique implementation of any SCF in the information elicitation problem.

Proposition 1. *In the information elicitation problem, any SCF f is uniquely implementable if equality (3), that is, NCKS, holds.*

4.1.1. Mechanism design

To prove Proposition 1, we design the following mechanism: $G=(M,g,x)$. Let[12]

$$K=1,$$

and

$$M_i=M_i^1=\Delta(\Omega)\text{ for all }i\in N.$$

We specify an allocation rule g as a majority rule: for each $m\in M$,

$$g(m)=f(\omega)\quad \text{if } m_i^1=\omega \text{ for more than } n/2 \text{ agents},$$

and

$$g(m)=a^*\quad \text{if there exists no such } \omega,$$

where a^* is selected arbitrarily. For each $i\in N$ and $j\neq i$, we specify $y_{i,j}:M_i^1\times M_j^1\to[-1,0]$ as a quadratic scoring rule:

[12] For the information elicitation problem alone, mechanism design with $K=1$ is enough. We will need $K\geq 2$ for general implementation problems, because we utilize the Abreu-Matsushima method. For its detail, see Subsection 4.2.

$$y_{i,j}(m_i^1, m_j^1) = -\sum_{\omega\in\Omega}\{m_i^1(\omega) - m_j^1(\omega)\}^2,$$

which implies the distance between agent i's 1-st sub-message and agent j's 1-st sub-message. For each $i \in N$, we specify the payment rule: for every $m \in M$,

$$\hat{x}_i(m) = \frac{\varepsilon}{n-1}\sum_{j\neq i} y_{i,j}(m_i^1, m_j^1)$$
$$= -\frac{\varepsilon}{n-1}\sum_{j\neq i}[\sum_{\omega\in\Omega}\{m_i^1(\omega) - m_j^1(\omega)\}^2],$$

where we denote $m_i = m_i^1$ because of $K = 1$.

From the nature of the quadratic scoring rule, any selfish type prefers to mimic the average of the other agents' messages. Importantly, any honest type prefers announcing slightly more honestly than selfish types. These are the driving forces that tempt even selfish types to announce truthfully.

4.1.2. Example

The following characteristics of the quadratic scoring rule are crucial for understanding Proposition 1. For simplicity of the arguments, this subsection focuses on the two-agent case.

(a) Each agent's message space is not the set of states but the set of probability distributions over states. Hence, an agent can continuously fine-tune their message and payment.
(b) Any selfish agent is incentivized to match her message with the other agent's message.
(c) Any honest agent is also incentivized to match her message with the other agent's message, but due to the intrinsic preference for honesty, she wants to behave slightly more honestly than the other agent.
(d) Suppose that agent 1 expects the possibility that agent 2 makes an announcement more honestly than what agent 2 expects about agent 1's announcement. Then, since agent 1 rationally expects that agent 2 attempts to announce more honestly, agent 1 with selfish type has an incentive to make the announcement more honestly than agent 2 initially expects. The same scenario holds even if agents 1 and 2 are replaced. This will be the driving force for a tail-chasing competition through which each agent announces more honestly than the other, reaching honest reporting by both.

Whenever agent i expects the possibility that the other agent $j \neq i$ is honest, then the supposition in (d) holds and agent i is driven to be more honest. However, the other agent j does not have to be honest: it is necessary and sufficient that agent i expects the possibility that the other agent j, whether selfish or honest, is driven to be more honest.

Consider the following example with a binary state space and a finite type space, where $n = 2$, $\Omega = \{0, 1\}$, and $T_i = \{1, ..., H\}$ for each $i \in \{1, 2\}$. We assume that agent i is honest if and only if $t_i = 1$, that is, $E_i^* = \{2, ..., H\}$. The message space of agent i is given by $M_i = [0, 1]$, where $m_i \in [0, 1]$ indicates the probability that state 1 ($\omega = 1$) occurs. The quadratic scoring rule is given by

$$\hat{x}_1(m) = \hat{x}_2(m) = -(m_1 - m_2)^2.$$

We assume that the common prior is symmetric; that is, $\pi(h, h') = \pi(h', h)$ for all $(h, h') \in \{1, ..., H\}^2$. Because the mechanism and agents are symmetric, we often refer to an agent with type h as a "type-h agent" without specifying their identity $i \in \{1, 2\}$. We assume that the set of selfish types $E_i^* = \{2, \ldots, H\}$ is path-connected in the sense that

$$\pi(h, h+1) > 0 \text{ for all } h \in \{2, \ldots, H-1\}.$$

Without loss of generality, we assume that the true state is $\omega = 1$ (the analysis for the case of $\omega = 0$ is similar), and we drop it from the notation. The psychological cost for each agent i with honest type is given by $\lambda(1 - m_i)$, where $\lambda > 0$. Let $\bar{m}_j(t_i; s_j)$ be agent j's expected message conditional on agent i's type t_i:

$$\bar{m}_j(t_i; s_j) \equiv E\left[s_j(t_j)|t_i\right] = \sum_{h=1}^{H}\pi_i(h|t_i)s_j(h).$$

Then, agent i's best response against s_j is given by

$$BR_i(s_{-i}, t_i) = \bar{m}_j(t_i; s_j) \quad \text{if } t_i \in \{2, \ldots, H\},$$

and

$$BR_i(s_{-i}, t_i) = \min\left\{\bar{m}_j(t_i; s_j) + \frac{\lambda}{2}, 1\right\} \quad \text{if } t_i = 1.$$

Hence, any honest agent is driven to be more honest than a selfish agent.

Case 1. First, consider the case in which the set of selfish types is disconnected from the honest type, that is,

$$\pi(1,h)=0 \text{ for all } h\in\{2,\dots,H\}.$$

Any selfish agent expects that the other agent is selfish with certainty, and any honest agent expects that the other agent is honest with certainty.

When $t=(1,1)$ is realized, the best response of each (honest) agent $i\in\{1,2\}$ is given by $s_i(1)=\min\{s_j(1)+\lambda/2,1\}$; that is, the preference for honesty drives each agent to select a message that is slightly more honest than the other. Clearly, whenever s is a BNE, $s_1(1)=s_2(1)=1$ must be satisfied.

In contrast, an equilibrium strategy can take any value when $h\in\{2,\dots,H\}$. As long as there exists a constant $p\in[0,1]$ such that

$$s_i(h)=p \text{ for all } i\in N \text{ and } h\in\{2,\dots,H\},$$

it is a BNE. Hence, there are infinitely many BNEs in which any selfish agent may tell a lie. Clearly, we fail to elicit the correct state as a unique BNE in Case 1.

Case 2. Consider the case in which, unlike Case 1, the set of selfish types is connected with the honest type in a minimal sense such that there exists $h\in\{2,\dots,H\}$ with $\pi(1,h)>0$. For simplicity, we assume that $h=2$, that is,

$$\pi(1,2)>0.$$

It is easy to see that the same argument holds even if we replace type 2 with any $h\in\{3,\dots,H\}$.

Because of higher-order reasoning, this minimal connection drastically changes the set of BNEs as follows. Clearly, a type-1 (honest) agent is driven to be more honest. The minimal connection implies that a type-2 agent expects that the other agent may be type-1 with a positive probability. Since a type-2 agent would like to match her message with the other agent (who could be type-1), she is also driven to be more honest. Similarly, a type-3 agent expects that the other agent may be type-2 with a positive probability and, thus, is driven to be more honest. We can iterate this argument and verify that any agent, whether selfish or honest, is driven to be more honest, that is, attempts to send a more honest message than the other. This structure of best responses immediately leads us to the uniqueness of BNE, where all agents report truthfully.

Note that this uniqueness holds even if both agents' selfishness is mutual knowledge. As long as $t_1\geq 3$ and $t_2\geq 3$, each agent does not expect that the other agent may be honest. However, the aforementioned higher-order reasoning will guide any agent to send a more truthful message, which drastically shrinks the set of BNE. Under NCKS, this logic generally functions and the uniqueness of the BNE is guaranteed.

Case 1 corresponds to situations in which all selfish types completely eliminate associations with honest types. In this case, unique information elicitation is impossible. By contrast, as in Case 2, if there is at least one selfish type who expects even a little (possibly indirect) influence of an honest type, then unique information elicitation is achievable. The driving force behind this phenomenon is not that more people become honest but that selfish people do not rule out the existence of honest agents from their epistemological considerations.

4.1.3. Proof of Proposition 1

From the nature of the quadratic scoring rules, it is clear that the sincere strategy profile s^* is a BNE in the game associated with the specified mechanism G; thus, it suffices to show uniqueness.

Suppose that s is a BNE. Fix $\omega\in\Omega$ arbitrarily. Let

$$\alpha=\min_{(i,t_i)} s_i(\omega,t_i)(\omega),$$

and

$$\tilde{T}_i\equiv\{t_i\in T_i | s_i(\omega,t_i)(\omega)=\alpha\} \text{ for each } i\in N.$$

Suppose that NCKS, that is, the equality (3), holds. Note that NCKS is equivalent to

$$V_i^{\infty}(E^*)=\phi \text{ for all } i\in N.$$

Suppose that $\alpha<1$, that is, there exists an agent $i\in N$ with type $t_i\in T_i$ that does not adopt the sincere strategy. Note from the definition of psychological cost that any honest agent prefers making announcements more honestly than selfish agents. Hence, no honest type belongs to $\tilde{T}_i$; that is, $\tilde{T}_i\subset E_i^*$.

Let us consider an arbitrary $i \in N$ and $t_i \in \tilde{T}_i$. Note that α equals the average of the other agents' announcements on ω in expectation but not greater than any announcement. Hence, an epistemological type t_i assumes that any other agent $j \neq i$ announces $m_i(\omega) = \alpha$, that is,

$$\pi_i \left(\underset{j \in N}{\times} \tilde{T}_j \middle| t_i \right) = 1.$$

This, along with the definition of the psychological cost, implies that agents i with epistemological type t_i expect that the other agents are surely selfish, that is,

$$\pi_i(E^*|t_i) = 1.$$

Hence, we have

$$\tilde{T}_i \subset V_i^1(E^*).$$

Moreover, since

$$\pi_i \left(\underset{j \in N}{\times} V_j^1(E^*) \middle| t_i \right) \geq \pi_i \left(\underset{j \in N}{\times} \tilde{T}_j \middle| t_i \right) = 1,$$

we have $\pi_i(\times_{j \in N} V_j^1(E^*)|t_i) = 1$, that is,

$$\tilde{T}_i \subset V_i^2(E^*).$$

Similarly, we have

$$\tilde{T}_i \subset V_i^k(E^*) \text{ for all } k \geq 2.$$

Hence, we have

$$\tilde{T}_i \subset V_i^\infty(E^*),$$

which contradicts the assumption that $V_i^\infty(E^*) = \phi$. Hence, we conclude that $\alpha = 1$, or, equivalently, $s_i(\omega, t_i) = \omega$ for all $\omega \in \Omega$. Accordingly, $s = s^*$ must be the case for any BNE. From these observations, we prove Proposition 1.

Remark 1. It is well known that quadratic scoring rules incentivize agents to be honest as a BNE in the information elicitation problem. The seminal work of Matsushima and Noda (2020) first pointed out that truth-telling is not only a BNE but also a unique BNE, provided that "all agents are selfish" is not mutual knowledge. Proposition 1 applies the logic of higher-order beliefs such as email games (Rubinstein, 1989) and global games (Carlsson and van Damme, 1993; Morris and Shin, 1998) to the information elicitation problem, and succeeded in extending the findings of Matsushima and Noda (2020) to the situation in which "all agents are selfish" may be mutual knowledge but is not common knowledge.

4.2. General case

Proposition 1 depends on the assumption that each agent's material payoff is irrelevant to the allocation. This subsection eliminates this assumption and completes the full proof of Theorem 1. By integrating the tail-chasing method of quadratic scoring rules into another tail-chasing method originated in the Abreu-Matsushima mechanism (Abreu and Matsushima, 1992a, 1992b), we show that any SCF is uniquely implementable even under severely limited liability.

4.2.1. Mechanism design

To prove Theorem 1 generally, we design another mechanism G as follows. Let

$$K \geq 3,$$
$$M_i^1 = \Delta(\Omega),$$

and

$$M_i^k = \Omega \text{ for all } k \in \{2, ..., K\}.$$

At the first sub-message, each agent i announces a probability distribution over states. At each sub-message other than the first, she announces a degenerate distribution, that is, a single state. For each $k \in \{3, ..., K\}$, we specify $g^k : M^k \to \Delta(A)$ as a majority rule: for every $\omega \in \Omega$,

$$g^k(m^k) = f(\omega) \quad \text{if } m_i^k = \omega \text{ for more than } n/2 \text{ agents},$$

and

$$g^k(m^k) = a^* \quad \text{if there exists no such } \omega.$$

The central planner randomly selects $k \in \{3, ..., K\}$ and determines the allocation according to $g^k(m^k) \in \Delta(A)$; that is, we specify the allocation rule g:

$$g(m) = \frac{\sum_{k=3}^{K} g^k(m^k)}{K-2} \text{ for all } m \in M.$$

It is important to note that $g(m)$ is independent of the 1-st and 2-nd sub-messages, m^1 and m^2.

To specify the payment rule, we use the quadratic scoring rules $y_{i,j}$ as well as the following functions, z_i, w_i, and r_i; that is, we define $z_i : M_i^2 \times M_{i+1}^1 \to [-1, 0]$:

$$z_i(m_i^2, m_{i+1}^1) = -1 \quad \text{if } m_i^2 \neq m_{i+1}^1,$$

and[13]

$$z_i(m_i^2, m_{i+1}^1) = 0 \quad \text{if } m_i^2 = m_{i+1}^1,$$

which implies that the agent i is fined if her 2-nd sub-message is different from her neighbor's 1-st sub-message. Importantly, we define $w_i : M \to [-1, 0]$:

$$w_i(m) = -1 \quad \text{if there exists } k \in \{3, ..., K\} \text{ such that } m_i^k \neq m_i^2, \text{ and } m_j^{k'} = m_j^2 \text{ for all } k' < k \text{ and } j \in N,$$

and

$$w_i(m) = 0 \quad \text{if there exists no such } k \in \{3, ..., K\},$$

which implies that agent i is fined if she is the first deviant from the own 2-nd sub-message. The function w_i plays a central role in creating a tail-chasing method originated in Abreu and Matsushima (1992a, 1992b), which incentivizes agents to make the truthful announcements at all sub-messages after the third as unique equilibrium behavior. We further define $r_i(m_i) \in \{0, ..., K-2\}$ as the number of integers $k \in \{3, ..., K\}$ such that $m_i^k \neq m_i^2$, that is, the number of agent $i's$ sub-messages after her 3-rd sub-message that are different from her 2-nd sub-message.

Fix an arbitrary positive real number $\xi > 0$, which is set sufficiently large. We specify the payment rule x_i for agent i:

$$x_i(m) = \frac{\varepsilon}{3+\xi}\{\frac{1}{n-1}\sum_{j \neq i} y_{i,j}(m_i^1, m_j^1) + \xi z_i(m_i^2, m_{i+1}^1) + w_i(m) - \frac{r_i(m_i)}{K-2}\}.$$

Note that the specified payment rule x satisfies limited liability; that is,

$$x_i(m) \in [-\varepsilon, \varepsilon] \text{ for all } i \in N \text{ and } m \in M.$$

Let us select $K \geq 3$ sufficiently large to satisfy

$$K > \frac{3+\xi}{\varepsilon} \max_{(a,a') \in A^2, i \in N} \{v_i(a, \omega) - v_i(a', \omega)\} + 2. \tag{4}$$

With $n \geq 3$, the sincere strategy profile s^* satisfies that for every $(\omega, t) \in \Omega \times T$, $m \in M$, and $i \in N$,[14]

$$g^k(m^k) = f(\omega) \quad \text{if } m_{-i} = s_{-i}^*(\omega, t_{-i}),$$
$$x_i(m) = 0 \quad \text{if } m = s^*(\omega, t),$$

and

$$x_i(m) < 0 \quad \text{if } m_{-i} = s_{-i}^*(\omega, t_{-i}) \text{ and } m_i \neq s_i^*(\omega, t_i).$$

This implies that s^* is a BNE, and it achieves the value of the SCF f without monetary transfers on the equilibrium path.

The next subsection will show that if a strategy profile s is a BNE in the game associated with the specified mechanism, then $s = s^*$ must hold, which completes the proof of Theorem 1.

[13] We denote $i + 1 = 1$ if $i = n$.

[14] We denote $s_{-i}(\omega, t_{-i}) = (s_j(\omega, t_j))_{j \neq i}$.

4.2.2. Proof of Theorem 1

The proof of Theorem 1 is divided into two parts: "*information elicitation*" and "*implementation with provability*" in the following manner.

Part 1 (Information Elicitation): Part 1 shows that $s^1 = s^{*1}$, that is, every agent, whether selfish or honest, announces the state truthfully for the 1-st sub-message. Note that each agent $i's$ 1-st sub-message influences her welfare only through $\sum_{j\neq i} y_{i,j}(m_i^1, m_j^1)$. Hence, we can directly apply Proposition 1 and show that $s^1 = s^{*1}$.

Part 2 (Implementation with Provability): Assume that a BNE strategy profile s satisfies $s^1 = s^{*1}$. Part 2 shows that $s^k = s^{*k}$ for all $k \in \{2, ..., K\}$, that is, all agents announce the state truthfully at their remaining sub-messages.

First, because $\xi z_i(m_i^2, m_{i+1}^1)$ imposes a relatively large fine, each agent i is willing to announce truthfully for the 2-nd sub-message, irrespective of the other sub-messages of this agent.

Each agent i regards her 2-nd sub-message as reference, and she is tempted to announce this reference at any sub-message $k \in \{3, ..., K\}$ of this agent. Given that this reference is equivalent to the true state in equilibrium, that is, the state that actually occurs is substantially provable, it follows that all agents are tempted to announce the state truthfully at every sub-message.

To understand the logic behind Part 2, consider a case in which ε is sufficiently large to satisfy

$$\frac{\varepsilon}{3+\xi} > \max_{(a,a')\in A^2} \{v_i(a,\omega) - v_i(a',\omega)\}. \tag{5}$$

From (4) and (5), we can select $K = 3$ and simply write the designed mechanism as follows: for every $\omega \in \Omega$ and $m \in M$ such that $m_i^1 = m_i^2 = \omega$ for all $i \in N$,

$$g(m) = f(\tilde{\omega}) \quad \text{if } m_i^3 = \omega \text{ for more than } n/2 \text{ agents},$$
$$g(m) = a^* \quad \text{if there exists no such } \tilde{\omega},$$

and

$$x_i(s_i^*(\omega, t_i), m_{-i}) - x_i(m) = \frac{2\varepsilon}{3+\xi} \geq \frac{\varepsilon}{3+\xi} \quad \text{if } m_i^3 \neq s_i^{*3}(\omega, t_i) = \omega.$$

From (5), we have

$$x_i(s_i^*(\omega, t_i), m_{-i}) - x_i(m) > \max_{(a,a')\in A^2} \{v_i(a,\omega) - v_i(a',\omega)\}$$
$$\geq v_i(g(s_i^*(\omega, t), m_{-i}), \omega) - v_i(g(m), \omega).$$

Hence, the penalty on lying for the 3-rd sub-message is greater than the impact of this on the determination of allocation, and $s^3 = s^{*3}$ must hold.

Following Abreu and Matsushima (1992a), we can extend this observation to the case where ε is small, if we select a sufficiently large K to satisfy (4). The designed mechanism incentivizes each agent to avoid being the first deviant starting from the 3-rd sub-message and also provides each agent i with an incentive to reduce the number $r_i(m_i)$. This method drives all agents into a tail-chasing competition toward honest reporting from the 3-rd to the K-th sub-messages. Hence, $s = s^*$ must hold.

The formal proof of Theorem 1 is as follows. Suppose that s is a BNE. Fix an arbitrary state $\omega \in \Omega$. First, we show that

$$s_i^1(\omega, t_i) = \omega \text{ for all } i \in N \text{ and } t_i \in T_i.$$

Because the selection of m_i^1 influences agent $i's$ welfare only through $\sum_{j\neq i} y_{i,j}(m_i^1, m_j^1)$ and the psychological cost, the following properties are obtained:

$$[\theta_i(t_i) = 0]$$
$$\Rightarrow [s_i^0(\omega, t_i) \in \arg\max_{m_i\in M_i} E[\sum_{j\neq i} y_{i,j}(m_i^1, m_j^1) | \omega, t_i, s_{-i}, G]],$$

and

$$[\theta_i(t_i) = 1] \Rightarrow [s_i(\omega, t_i) \in \arg\max_{m_i\in M_i} E[\sum_{j\neq i} y_{i,j}(m_i^1, m_j^1)$$
$$- c_i(m_i, \omega, t_i) | \omega, t_i, s_{-i}, G].$$

From the nature of the quadratic scoring rule and the psychological cost, we can calculate the best response as follows:

$$[\theta_i(t_i)=0] \Rightarrow [s_i^1(\omega,t_i)=E[\frac{\sum_{j\neq i} s_i^1(\omega,t_j)}{n-1}|\omega,t_i]],$$

and

$$[\theta_i(t_i)=1] \Rightarrow [\text{either } s_i^1(\omega,t_i)(\omega)=1 \text{ or}$$
$$s_i^1(\omega,t_i)(\omega) > E[\frac{\sum_{j\neq i} s_i^1(\omega,t_j)(\omega)}{n-1}|\omega,t_i]].$$

In other words, any selfish agent mimics the average of the other agents' 1-st sub-messages in expectation, while any honest agent announces more honestly than the selfish types. This will drive agents into a tail-chasing competition, reaching the point at which all agents report honestly at their 1-st sub-messages. Hence, from Proposition 1, we can prove that any BNE s satisfies $s_i^1 = s_i^{*1}$ for all $i \in N$.

Because $\xi z_i(m_i^2, m_{i+1}^1)$ imposes a relatively large fine, it follows from $s_{i+1}^1 = s_{i+1}^{*1}$ that each agent i is willing to select $m_i^2 = \omega$. Hence, $s_i^2 = s_i^{*2}$ must hold for all $i \in N$.

We further prove that

$$s_i^k(\omega,t_i)=\omega \text{ for all } k\in\{3,...,K\},\ i\in N,\ \text{and } t_i\in T_i.$$

The specifications of w_i and r_i imply that if an agent i announces a sub-message different from her 2-nd sub-message as the first deviation starting from the 3-rd sub-message, she is fined the monetary amount $\frac{\varepsilon}{3+\xi}$. Because we have selected a sufficiently large K, that is, inequality (4) holds, the impact of the selection of each sub-message on the determination of the allocation is sufficiently small compared with the monetary amount $\frac{\varepsilon}{3+\xi}$. This will drive agents into another tail-chasing competition, originated in Abreu and Matsushima (1992a, 1992b), through which each agent avoids becoming the first deviant. Because we have already proved that all agents announce truthfully at their 2-nd sub-messages, this competition drives them to announce the state truthfully at all sub-messages.

To be precise, consider an arbitrary $k \in \{3,...,K\}$ and suppose that $s^{k'} = s^{*k'}$ for all $k' < k$. If $m_j^k \neq \omega$ for some $j \neq i$, agent i strictly prefers announcing truthfully at the k-th sub-message, because she can avoid being the first deviant. Even if $m_j^k = \omega$ for all $j \neq i$, agent i still strictly prefers announcing truthfully at the k-th sub-message because she does not want to increase $r_i(m_i)$. Hence, through the iterative elimination of dominated strategies, we can inductively prove that $s_i^k = s_i^{*k}$ for all $i \in N$ and $k \in \{3,...,K\}$. In other words, there exists no BNE other than the sincere strategy profile s^*.

Because s^* is a BNE and achieves the value of f, we have completed the proof of Theorem 1.

Remark 2. The specified payment rule consists of four components, $y_{i,j}$, z_i, w_i, and r_i. The first component $y_{i,j}$ corresponds to the quadratic scoring rules that solve Part 1 (Information Elicitation). To separate Part 1 and Part 2 (Implementation with Provability), we use the second sub-messages as references rather than the first sub-messages. The second component z_i serves as an incentive scheme to match the second sub-messages (the references) with the first sub-messages (the true state). The third component w_i corresponds to the logical core of the Abreu-Matsushima mechanism that makes all sub-messages after the third truthful through a tail-chasing procedure in which all agents dislike to be the first deviants from the references. The fourth component r_i plays an auxiliary role in this procedure, which helps to eliminate white lies and make the best responses unique. Finally, the inequality of (4) ensures that these four components function consistently with each other.

Remark 3. To better understand mechanism G, let us consider the following decision procedure: The central planner asks each agent to input a distribution on Ω. The central planner also asks each agent to input an element of Ω, and gives all agents a one-hour grace period. During this continuous period, each agent can change her second input at any time and number of times. After this grace period, the first and second inputs become public. The central planner selects one point from the grace period and then determines the allocation according to the majority rule and their inputs at this point. The central planner imposes just a small monetary fine to any agent who is the last person to change the second input. The central planner imposes another small, but slightly larger, monetary fine to any agent whose initial second input is different from her neighbor's first input. The central planner also makes monetary transfers according to the quadratic scoring rule and the first inputs.

Because of the nature of the quadratic scoring rule, all agents are willing to make their first inputs equivalent to the true state. Given that the possibility of selecting a point in the continuous period is negligible, any agent prefers to make her initial input equal to the true state and avoid becoming the last person to change the second input: she is willing to keep her correct input unchanged during the grace period.

Remark 4. In the proof of Theorem 1, we have proved the uniqueness of not only the pure but also the mixed-strategy BNE. In fact, in Part 1, any agent has the unique best response of the 1-st sub-message to any mixture of the other agents' 1-st

sub-messages. In Part 2, we eliminated all unwanted strategies through an iterative dominance process. This guarantees the uniqueness of the mixed-strategy BNE.

Remark 5. Without substantial difficulty, we can replace the payment rule with a budget-balancing payment rule. We redefine w_i by replacing the first deviant among all agents with the first deviant among all agents other than some agent and set it as the transfer from this agent. We also set other parts of the payment rule as the transfers from agents whose announcements are irrelevant.

5. Conclusion

This study investigated a society in which people are either selfish or honest, and showed that every incentive-compatible SCF, whether material or nonmaterial, is implementable in BNE if "all agents are selfish" never happens to be common knowledge.

This study assumed that there exist only two motives for agents: selfishness and honesty. In reality, there could be adversarial motives such as "always tell a lie." However, our results are robust to the consideration of such motives although it is not explicitly shown in this study. The equilibrium messages are attracted to somewhere close to truth-telling whenever these motives are not as important as honesty; that is, the central planner can still identify the true state by checking whether agents' messages are attracted by a certain message through the indications.

Our findings will bring hope to central planners who lack the information necessary for normative judgments such as "Are social benefits fairly distributed in the society?", "Who needs relief from poverty?", "How will decision-making affect outsiders and future generations?", and others. Selfish people are generally unmotivated by such ethical concerns even if they have a keen interest in ethical concerns and are knowledgeable about them. From the viewpoint of social network (Putnam, 2006) and social common capital (Uzawa, 2005) in epistemology, the common knowledge assumption on selfishness implies that society is divided into a group of selfish people and a group of honest people and these groups are disconnected from each other. With this common knowledge, the central planner cannot derive ethical information from selfish people correctly. However, if selfish people and honest people are path-connected with each other, the central planner can properly derive such information even from selfish agents and reflect it in her normative judgment, that is, she can implement any ethical SCF she desires.

This study considered the role of a social network epistemologically, implicitly assuming that preference for honesty and prosocial propensity are consistent. Consideration of the situation where this premise does not hold is an important issue. The situation where the SCF is given for the central planner's private purpose, which agents do not agree with, is an example. Matsushima (2013) analyzes the unique implementation in this situation as the possibility of psychological guidance, which causes those agents to reveal information truthfully through a manipulation of the revelation process. Another example is a situation such as scarce resource allocation in a pandemic, where agents have different ethical criteria and the central planner composes an SCF by finding those compromises. The related works are Pathak et al. (2020) and Matsushima (2021b, 2021c). Such research is expected to develop further in the future.

Declaration of competing interest

The author has no conflicts of interest directly relevant to the content of this article.

References

Abeler, J., Nosenzo, D., Raymond, C., 2019. Preference for truth-telling. Econometrica 87 (4), 1115–1153.
Abreu, D., Matsushima, H., 1992a. Virtual implementation in iteratively undominated strategies: complete information. Econometrica 60, 993–1008.
Abreu, D., Matsushima, H., 1992b. Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information. Mimeo. https://www.econexp.org/hitoshi/AMincomplete.pdf.
Aoyagi, M., 1998. Correlated types and Bayesian incentive compatible mechanisms with budget balance. J. Econ. Theory 79, 142–151.
Arrow, K., 1951. Social Choice and Individual Values. Yale University Press, New Haven.
Brier, G., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3.
Carlsson, H., van Damme, E., 1993. Global games and equilibrium selection. Econometrica 61, 989–1018.
Cooke, R., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press, New York.
Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 76 (6), 1579–1601.
Dasgupta, A., Ghosh, A., 2013. Crowdsourced judgement elicitation with endogenous proficiency. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 319–330.
Dogan, B., 2017. Eliciting the socially optimal allocation from responsible agents. J. Math. Econ. 73, 103–110.
Dutta, B., Sen, A., 2012. Nash implementation with partially honest individuals. Games Econ. Behav. 74 (1), 154–169.
Ellingsen, T., Johannesson, M., 2004. Promises, threats and fairness. Econ. J. 114 (495), 397–420.
Gibbard, A., 1973. Manipulation of voting schemes: a general result. Econometrica 41 (4), 587–601.
Hurwicz, L., 1972. On informationally decentralized systems. In: McGuire, C.B., Radner, R. (Eds.), Decision and Organization. North-Holland, Amsterdam.
Jackson, M., 2001. A crash course in implementation theory. Soc. Choice Welf. 18, 655–708.
Johnson, S., Pratt, J., Zeckhauser, R., 1990. Efficiency despite mutually payoff-relevant private information: the finite case. Econometrica 58, 873–900.
Kartik, N., 2009. Strategic communication with lying costs. Rev. Econ. Stud. 76 (4), 1359–1395.
Kartik, N., Ottaviani, M., Squintani, F., 2007. Credulity, lies, and costly talk. J. Econ. Theory 134 (1), 93–116.
Kartik, N., Tercieux, O., 2012. Implementation with evidence. Theor. Econ. 7 (2), 323–355.

Kartik, N., Tercieux, O., Holden, R., 2014. Simple mechanisms and preferences for honesty. Games Econ. Behav. 83, 284–290.
Kong, Y., Schoenebeck, G., 2019. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. ACM Trans. Econ. Comput. 7 (1), 1–33.
Kreps, D.M., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated prisoners' dilemma. J. Econ. Theory 27, 245–252.
Lombardi, M., Yoshihara, N., 2017. Natural implementation with semi-responsible agents in pure exchange economies. Int. J. Game Theory 46 (4), 1015–1036.
Lombardi, M., Yoshihara, N., 2018. Treading a fine line: (im)possibilities for Nash implementation with partially-honest individuals. Games Econ. Behav. 111, 203–216.
Lombardi, M., Yoshihara, N., 2019. Partially-honest Nash implementation: a full characterization. Econ. Theory 54 (1), 1–34.
Maskin, E., 1977/1999. Nash equilibrium and welfare optimality. Rev. Econ. Stud. 66, 23–38.
Maskin, E., Sjöström, T., 2002. Implementation theory. In: Arrow, K., Sen, A., Suzumura, K. (Eds.), Handbook of Social Choice and Welfare, vol. 1. Elsevier.
Matsushima, H., 1990. Dominant strategy mechanisms with mutually payoff-relevant information and with public information. Econ. Lett. 34, 109–112.
Matsushima, H., 1991. Incentive compatible mechanisms with full transferability. J. Econ. Theory 54, 198–203.
Matsushima, H., 1993. Bayesian monotonicity with side payments. J. Econ. Theory 59, 107–121.
Matsushima, H., 2007. Mechanism design with side payments: individual rationality and iterative dominance. J. Econ. Theory 133 (1), 1–30.
Matsushima, H., 2008a. Behavioral aspects of implementation theory. Econ. Lett. 100 (1), 161–164.
Matsushima, H., 2008b. Role of honesty in full implementation. J. Econ. Theory 139, 353–359.
Matsushima, H., 2013. Process manipulation in unique implementation. Soc. Choice Welf. 41 (4), 883–893.
Matsushima, H., 2019. Implementation without expected utility: ex-post verifiability. Soc. Choice Welf. 53 (4), 575–585.
Matsushima, H., 2021a. Partial ex-post verifiability and unique implementation of social choice functions. Soc. Choice Welf. 56, 549–567.
Matsushima, H., 2021b. Assignments with Ethical Concerns. Discussion Paper CARF-F-514 (UTMD-007). University of Tokyo.
Matsushima, H., 2021c. Auctions with Ethical Concerns. Discussion Paper CARF-F-515 (UTMD-008). University of Tokyo.
Matsushima, H., 2021d. Epistemological Implementation of Social Choice Functions. Discussion Paper CARF-F-518 (UTMD-013). University of Tokyo.
Matsushima, H., Noda, S., 2020. Mechanism Design with Blockchain Enforcement. CARF-F-474. University of Tokyo.
Mazar, N., Amir, O., Ariely, D., 2008. More ways to cheat — expanding the scope of dishonesty. J. Mark. Res. 45 (6), 651–653.
Miller, N., Pratt, J., Zeckhauser, R., Johnson, S., 2007. Mechanism design with multidimensional, continuous types and interdependent valuations. J. Econ. Theory 136 (1), 476–496.
Miller, N., Resnick, P., Zeckhauser, R., 2005. Eliciting informative feedback: the peer-prediction method. Manag. Sci. 51 (9), 1359–1373.
Moore, J., 1992. Implementation in environments with complete information. In: Laffont, J.J. (Ed.), Advances in Economic Theory: Sixth World Congress. Cambridge University Press, Cambridge.
Morris, S., Shin, H.S., 1998. Unique equilibrium in a model of self-fulfilling currency attacks. Am. Econ. Rev. 88 (3), 587–597.
Mukherjee, S., Muto, N., Ramaekers, E., 2017. Implementation in undominated strategies with partially honest agents. Games Econ. Behav. 104, 613–631.
Ortner, J., 2015. Direct implementation with minimally honest individuals. Games Econ. Behav. 90, 1–16.
Pathak, P., Sonmez, T., Unver, U., Yenmez, M., 2020. Leaving no ethical value behind: triage protocol design for pandemic rationing. SSRN. https://ssrn.com/abstract=3569307.
Postlewaite, A., Vives, X., 1987. Bank runs as an equilibrium phenomenon. J. Polit. Econ. 95, 485–491.
Prelec, D., 2004. A Bayesian truth serum for subjective data. Science 306 (5695), 462–466.
Putnam, R., 2006. Bowling alone: Americas's declining social capital. J. Democr. 6 (1), 65–78.
Rubinstein, A., 1989. The electric mail game: strategic behavior under 'almost common knowledge'. Am. Econ. Rev. 79, 385–391.
Saporiti, A., 2014. Securely implementable social choice rules with partially honest agents. J. Econ. Theory 154, 216–228.
Satterthwaite, M., 1975. Strategy-proofness and arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions. J. Econ. Theory 10 (2), 187–217.
Savva, F., 2018. Strong implementation with partially honest individuals. J. Math. Econ. 78, 27–34.
Uzawa, H., 2005. Economic Analysis of Social Common Capital. Cambridge University Press, Cambridge, England.
Yadav, S., 2016. Selecting winners with partially honest jurors. Math. Soc. Sci. 83, 35–43.